[Classification: Social Sciences/Psychological and Cognitive Sciences]

Cross-language differences in the brain network subserving intelligible speech

Jianqiao Ge[a,b,1,2], Gang Peng[c,1], Bingjiang Lyu[b], Yi Wang[b], Yan Zhuo[d], Zhendong Niu[e], Li Hai Tan[f,2], Alexander P. Leff[g], Jia-Hong Gao[a,b,h,2]

[a]McGovern Institute for Brain Research, Peking University, Beijing, China

[b]Center for MRI Research, Peking University, Beijing, China

[c]Joint Research Centre for Language and Human Complexity, and Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong, China

[d]State Key Laboratory of Brain and Cognitive Science, Chinese Academy of Sciences, Beijing, China

[e]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

[f]Neuroimaging Unit, Department of Biomedical Engineering, School of Medicine, Shenzhen University, and Guangdong Key Laboratory of Biomedical Information Detection and Ultrasound Imaging, Shenzhen 518060, China

[g]Institute of Cognitive Neuroscience, University College London, London, UK

[h]Beijing City Key Lab for Medical Physics and Engineering, School of Physics, Peking University, Beijing, China

[1]These authors contributed equally to this work.

[2]Corresponding Author:

Jia-Hong Gao, Ph.D.

Center for MRI Research,

Peking University,

Beijing, China 100871.

Phone: +86-10-62753091

Fax: +86-10-62752918

E-mail: jgao@pku.edu.cn

Jianqiao Ge, Ph.D.

Center for MRI Research,

Peking University,

Beijing, China 100871

Phone: +86-10-62745073

Fax: +86-10-62752918

E-mail: gejq@pku.edu.cn

Li Hai Tan, Ph.D.

School of Medicine,

Shenzhen University,

E-mail: lihaitan@gmail.com

Number of Figures: 2

Number of Tables: 2

Supplemental Material: 1

Number of Pages: 31

Number of Words of Abstract: 250

Number of Words of Main Text: 3,824

Keywords: speech perception, tonal language, functional MRI, cortical dynamics

Abstract

How is language processed in the brain by native speakers of different languages? It is a hot debate between one brain for all languages and different brains for different languages that emphasize commonality or specificity respectively. We investigated the cortical dynamics involved in processing two very diverse languages, a tonal language (Mandarin Chinese) and a non-tonal language (English). We used functional MRI and dynamic causal modeling (DCM) analysis to compute and compare brain network models exhaustively with all possible connections among nodes of interested language regions in temporal and frontal cortex, and found that the information flow from posterior to anterior portions of the temporal cortex was commonly shared by Chinese and English speakers during the speech comprehension, whereas the inferior frontal gyrus received neural signals from the left posterior portion of temporal cortex in English speakers, and from bilateral anterior portion of temporal cortex in Chinese speakers. Our results revealed that although speech processing is largely carried out in the common left-hemisphere classical language areas (Broca's and Wernicke's areas) and anterior temporal cortex, speech comprehension across different language groups depends on how these brain regions interact with each other. Moreover, the right anterior temporal cortex which is crucial for tone processing is equally important as its left homologue, the left anterior temporal cortex, in modulating the cortical dynamics in tone language comprehension. The current study pinpoints the importance of bilateral anterior temporal cortex in language comprehension that are downplayed or even ignored by popular contemporary models of speech

comprehension.

Significance Statement

It is generally known that language processing is left-hemisphere dominant. However, whether the interactions among the typical left-hemispheric language regions differ across different languages is largely unknown. One of the best methods to address this question is modeling cortical interactions across language groups, but usually highly constrained by the model space with prior hypothesis due to massive computation demands. With cloud-computing, we used fMRI DCM analysis to compare more than 4000 models of cortical dynamics among critical language regions in temporal and frontal cortex, established the bias-free information flow maps that were commonly shared or specific for processing intelligible speech in Chinese and English, and also revealed the neural dynamics between left and right hemispheres in the Chinese speech comprehension.

Introduction

The brain of a newborn discriminates the various phonetic contrasts used in different languages (1) by recruiting distributed cortical regions (2); by 6-10 months it is preferentially tuned to the phonemes in native speech that they have been exposed to (3, 4). In adult humans, the key neural nodes that subserve speech comprehension are located in the superior temporal cortex (5, 6) and the inferior frontal cortex (7). Do these regions combine in different ways depending on the type of language that is being processed? Little is known about whether and how information flows among these critical language nodes in native speakers of different languages.

As one of the unique capacities of the human brains, the nature of compositional languages (8) and their neural mechanisms have been the interests of scientific research for decades. There are more than 7,000 different spoken languages in the world nowadays that all serve for communication. By exploring the brain basis of the universal properties and specificity of different languages, it helps address the essential problems like the constitution of the knowledge of language, as well as how it is acquired (9). While the traditional universal grammar theory suggests that linguistic ability manifests itself without being taught (10), recent connectionist theory with a neural network approach emphasizes that interactions among primary systems of neuronal processing units which support the language acquisition and use, and the weights of connections among these units are gradually changed during learning and thus highly constrained by the language specific features (9,11,12). Following the connectionism, the dual pathway model of language based on evidence

from neuroimaging studies is framed to interpret the neural basis of language comprehension and production (13, 14), especially in the speech comprehension.

Previous studies have shown that intelligible speech is processed hierarchically in human neocortex, with the anterior temporal (13, 15, 16) and frontal cortices (Broca's area) 'above' the posterior region of superior temporal sulcus/gyrus (pSTS/pSTG, core of Wernicke's area), which itself receives inputs from primary and secondary auditory cortices (17). The dorsal pathway starts from the posterior region of the temporal cortex, through the dorsal part to the temporal-parietal cortex and then reaches the frontal cortex for a sound-motor projection, whereas the ventral pathway through the ventral side of the temporal lobe to the anterior regions and reaches the frontal cortex for a sound-meaning mapping (7, 18).

However, the cortical regions that process speech are likely to be common across languages, but how these regions interact with each other in the cortical network may depend on the distinctive phonetic-linguistic characteristics in different languages, particularly in ones that vary to a large extent such as tonal and non-tonal languages. Cross-language research has made possible the comparison of neural pathways underlying different languages that vary in linguistic features. Previous studies investigated the behavioral consequence of processing different languages such as pitch accent processing (19), but little is know about the underlying brain connectivity. The current study aims to explore the dynamic neural networks of processing intelligible speech in two different languages with such a featured phonological-semantic variant, Mandarin Chinese and English. These two languages

are the most widely spoken languages in the world, but differ in several aspects, such as the use of lexical tones. In tonal languages like Mandarin Chinese, suprasegmental features, different pitch patterns, serve to distinguish lexical meaning whereas in non-tonal languages pitch changes are less complex and do not convey lexical information. Except the lexical tone, Mandarin Chinese also includes more homophones than English, which also makes the sound-meaning mapping dependent more on the context information during the speech, thus may raise higher demands on the ventral pathway of the related neural network. To test this hypothesis, we examined the cortical dynamics underlying speech comprehension for two groups of native speakers of different languages.

In the study, we used fMRI and dynamic causal modeling (DCM) (20) to first investigate the cortical dynamics among left pSTG, aSTG and IFG of native speakers in a tonal (i.e. Chinese) language compared with a non-tonal (i.e. English) language with identical experimental design. Thirty native speakers in Chinese and twenty-six native speakers in English with matched age, gender and all right-handedness were scanned while presented with intelligible and unintelligible speech of their native languages (either Mandarin Chinese or English) in blocks, spoken by a male and a female. Subjects were instructed only to judge the gender of the speakers. The data of native English speakers was reanalyzed from a previous study(21). The brain activation of the intelligible effect and the effective connectivity among the three left-hemisphere brain regions were analyzed for both language groups under identical procedures and then put together for comparison.

Results

In both groups the contrast of intelligible > unintelligible speech revealed significant neural activities in left anterior temporal lobe, supplementary motor area, postcentral gyrus and the pars triangularis of left inferior frontal gyrus (Fig. 1*A* and Table S1).

To establish the basic neural dynamic network of processing intelligible speech in Chinese and English, we first constructed dynamic models that consisted of the three shared left hemisphere brain regions which were engaged in processing speech in both languages: left aSTG (region A), left pSTG (region P), and left IFG (region F). We computed an exhaustive series of models, varying input site (7 families) with all possible patterns of connectivity among the three nodes (63 models per family), which generated a total of 441 alternative models for each subject. These models were estimated and the evidence for each was compared using a family-level random effect, Bayesian Model Selection analysis (22). A Bayesian Model Average (BMA) analysis was then performed to provide average connectivity values for each connection across all possible models in the model space for each subject. These values were entered into both within- and between-group analysis for individual connections using one- and two-sample *t*-test with FDR correction. The results showed that for both groups the auditory signals entered the neural network through pSTG node which is, by definition, the lowest of the three nodes in the cortical hierarchy. In terms of inter-regional connections, hearing intelligible speech increased

the strength of the ventral forward connection, pSTG-to-aSTG, in both groups. There were however, clear group differences for other connections; specifically, the English speakers had a significantly stronger dorsal forward connection from pSTG to IFG, whereas in the Chinese speakers the two connections emanating from the aSTG: a backward connection to pSTG and a lateral connection to IFG were stronger. (Fig. 2*A* and Table 1).

In addition to the three shared brain regions in left hemisphere, the Chinese speakers had an additional activation in the right anterior temporal pole ($F_{1,51}$=8.141, P=0.006, Figs. 1*B* and 1*C*) during the processing of intelligible speech, consistent with previous findings that the anterior region of the right temporal cortex is functionally linked with pitch and tone processing (23). To investigate the comprehensive neural dynamics for the tonal language, we carried out a second analysis of the Chinese-only data which included this fourth region (right aSTP/STG). BMA analysis of an exhaustive set of 4095 alternative models was conducted (input into left pSTG, 12 connections systematically varied across models). The connections significantly modulated within the left hemisphere were revealed to be the same as those identified in the three region analysis (Fig. 2*B*). This analysis also identified three inter-hemispheric connections significantly modulated by intelligibility: the bidirectional right aSTG- left pSTG connections (same connectivity pattern as with the left aSTG-pSTG), and the right aSTG-to- left IFG connection (Table 2). Moreover, the individual modulation strength of intelligibility on connection of the left aSTG-to-left pSTG is positively correlated with the strength of the right aSTG-to-left

pSTG connection (Fig. 3*A*).

Discussion

This study found that during the speech comprehension processing, the three regions in the left hemisphere, inferior front gyrus (Broca's area), posterior temporal gyrus (Wernicke's area), and anterior temporal gyrus are shared by two language groups (both tonal and non-tonal), while the interactions among those regions depend on the language. Perceiving intelligible speech in a tonal language engages the bilateral ventral anterior temporal lobes and its connections with classical language areas are much stronger than in a non-tonal language, whereas the connection between the classical language areas in left hemisphere (from the posterior temporal lobe to inferior frontal lobe) is much stronger in a non-tonal (English) than in a tonal language (Chinese).

Importantly, both forward and backward connections from left and right aSTGs to left pSTS are involved in the tonal-language network. In tonal languages such as Chinese, suprasegmental features (e.g. pitch changes) are used to signify the meaning of a word, as well as having much greater homophony in daily vocabulary (24). The aSTG is considered to be a 'semantic hub' that it is critical in supporting language function (25), the underlying cortical pathway of speech processing based on these ventral connections are especially important to accomplish a more complicated sound-meaning mapping in a tonal language. We identified increased backward connections, which convey information about prior expectations in

hierarchical processing models (26), probably because the lack of suprasegmental phonological information initially, or that the auditory word pairs were heard in isolation, meaning that there was a lack of the usual phonology and sentence structure to help resolve word identity. The involvement of both temporal poles in this task may be due to either these increased task demands as homophony is much greater in Chinese than English, or the right hemisphere being more involved in processing pitch information in tonal languages (23, 27). The increased backward modulation from left anterior to the posterior parts of the temporal lobe was revealed to be significantly correlated with an increased modulation on the connection from right anterior part to the left posterior part of the temporal lobe, suggesting that the top-down modulation for further semantic processing on the ventral pathway is supported by integrated processing based on full phonological information including lexical tones from bilateral temporal lobes.

Moreover, the stronger forward connections between the anterior temporal poles and the Broca's area may be due to further semantic processing that is included in identifying word through the phonological information in Chinese. Subjects with greater difficulties in identifying the word (Fig. 3*B*) but intact performance in identifying the pronunciation (Fig. S4) showed greater connectivity on the bilateral connections from aSTG to the Broca's area (i.e. IFG). This preliminary result suggested an integrated forward processing of mapping the phonological information to the semantic-related representation on both hemispheres for a tonal language.

The only connection that was stronger in the English speakers was the forward

connection from pSTG to IFG. This likely represents a greater reliance of non-tonal languages on the dorsal stream, which is implicated in tasks that stress phonological (elemental speech-sound) processing of speech (28) where initial phonological features are informative enough for further communications in a non-tonal language.

The English group showed no activation in the right anterior temporal pole was probably because the absence of explicit demands on pitch-related processing in the experiment. Previous studies demonstrated that activation for speech processing depended on task (18,29). In tasks involving semantic, tone or intonation processing for English, significant activations could be detected in the right superior temporal area (30). In the current study, however, subjects were asked to merely make gender judgment on the auditory stimuli without any explicit requirement to understand the speech or make tone judgment, thus required minimal demand of tone-related processing for the intelligible speech. These results suggested an "automatic" engagement of the right temporal lobe for cortical dynamics in perceiving intelligible speech in a tonal language.

Our research has revealed for the first time that, particularly in tonal languages, classical left-hemisphere language areas such as Wernicke's and Broca's areas interact with the semantic system in anterior temporal lobes from both hemispheres when perceiving intelligible speech. At least two popular contemporary models of speech comprehension either downplay the importance (31) of these regions, or ignore them altogether (7). Note that the dynamic maps we describe here only reached to the sub-sentence level of speech comprehension, and further investigation is required to

consider sentence processing which is most frequently encountered in daily communication. As for the comparison between different language groups, the cross-center MRI data acquisition may induce potential confounding on images such as the distortion differences on the temporal lobe. Therefore, future cross-language comparison research on brain networks should consider these factors with extensive control on the data acquisition as well as language experience of the subjects.

Regardless of these limitations, our results advocate both language common and language specific cortical dynamics co-exist for speech comprehension, and emphasize the importance of the bilateral anterior temporal lobes and their connections with Wernicke's and Broca's areas in speech perception, particular for tonal languages such as Mandarin Chinese.

Materials and Methods

Subjects. Thirty native Chinese speakers and twenty-six native English speakers participated in the current research. The Chinese and English groups were matched on subjects' gender, ages, and handedness. Native English speakers participated in the previous study by Leff et.al (Details see Ref. 21) and the brain imaging data were reanalyzed in the current study. Native Chinese speakers participated in this study as paid volunteers (15 males, 15 females; aged between 21 and 28, mean age 24.2 years). All participants were right-handed, with normal hearing and normal or corrected-to-normal vision, Mandarin Chinese as their first language and had no neurological or psychiatric history. Written informed consent was obtained from each participant prior to scanning, and the study was conducted under the approval of local research ethics committee.

Stimuli. The experimental paradigm was adopted from a previous study about the cortical dynamics of intelligible English speech (21), while both intelligible and unintelligible auditory stimuli were presented to subjects to make gender judgment of the speakers. Half of the intelligible stimuli were idiomatic word pairs such as "cloud nine" and the other half were reordered word pairs of idioms (e.g. "mint nine"). To create their unintelligible counterparts, the stimuli of intelligible stimuli were time-reversed since this method removed the intelligibility of the forward speech while preserved the acoustic and voice identity information (5，21). In the current study, we designed the Chinese stimuli with consideration of matching the phonetic

and phonological characteristics of the idiomatic word pairs between Chinese and English. The English stimuli were mainly disyllabic. We selected Chinese intelligible stimuli with 3, 4 or 5-syllable in length that matched in duration with the English word-pairs. Half of the intelligible stimuli in Chinese were idiomatic words with 3-5 characters such as “ ” (he2 shi4 lao3, means “peacemaker”, the letters represent the official Romanization of standard Chinese, that is, Pinyin, while the number indicates the corresponding tone), “ ” (hua4 long2 dian3 jing1, means “finishing touch”), and the other half consisted of words from two unrelated idioms (e.g. “ ” (hong2 men2 hu4), “ ” (e4 guan4 hao4 long2); “ ” was the combination rearranged from the first two words of the idiom “ ”, e4 guan4 mang3 ying2 and the last two words of the idiom “ ”, ye4 gong1 hao4 long2). All intelligible stimuli were recorded digitally in a soundproof studio using Adobe Audition CS4 software. A male and a female speaker (both native speakers of Mandarin) produced all of the stimuli twice. As for Chinese unintelligible stimuli, the unintelligible counterparts of Chinese stimuli were also time-reversed of the intelligible stimuli that removed the intelligibility but preserved the acoustic and voice identity information. There were 84 auditory stimuli (half by male speaker, half by female speaker) for each of the three stimulus types, and no stimulus was repeated. All stimuli were edited for quality and length (885 ± 122 ms), and amplified so that there were no difference of loudness between speakers, as well as between the idioms and rearranged idioms.

Comparison and Psycholinguistic Analysis of Stimuli between Groups. We first calculated the duration of all auditory stimulus in English and Chinese group (English: 677-1080ms; Chinese: 730-1007 ms). A two-sample t-test revealed no significant difference on the duration between English and Chinese group. A similar procedure was also conducted to examine the stimulus loudness across languages and types of stimuli, and no significant difference was found. Both of the Chinese and English idioms were phrases established by usage and referred to a certain meaning, and the reordered idioms were created by combining two unrelated idiom. Therefore, the reordered idioms in both English and Chinese were meaningless but still recognizable syllable by syllable.

Since there are no comprehensive databases recording key psycholinguistic variables for either English or Chinese idioms, guided by the Cronk criteria (32), two psycholinguists who use English or Chinese as their native languages independently rated the idiomatic stimuli in their native language (either English or Chinese) on a binary scale, splitting the stimuli into the following, four uneven groups: 1) high familiarity, high literalness; 2) high familiarity, low literalness; 3) low familiarity, high literalness; 4) low familiarity, low literalness. Percentage proportions of stimuli that fell into each group were then calculated (Chinese: English): 1) 35%:33%; 2) 39%:31%; 3) 8%:11%; 4) 18%:35%. Non-parametric statistical tests of all four groups (independent samples Mann-Whitney U test) revealed no significant differences across the two languages (*P*s all = 1.0).

Procedure. In the experiment, subjects were required to make judgment about the gender of the speaker for each stimulus, and they were not informed of the types and contents of the stimuli in advance. Therefore the task was orthogonal to the effect of interest. The auditory stimuli were arranged in blocks by stimulus types with an alterable ratio of male and female speakers (2 : 5, 3 : 4, 4 : 3, or 5 : 2). All auditory stimuli only were presented once. There were twelve blocks for each type of stimuli, and they were evenly distributed in four scanning sessions. Each block was preceded by a preparation cue for 3150 ms, which was followed by seven trials. On each trial, an auditory stimulus was presented for 1180 ms, followed by a response cue for 2420 ms and then a fixation cross for 450 ms. The blocks were separated by a symbol "~" of 9450 ms presented at the center of the screen. Participants pressed one of the two buttons with their right index or middle finger to indicate their gender judgment of the speaker after the response cue immediately. The order of the blocks and the assignment of response buttons were counterbalanced across participants.

MRI Data Acquisition. Scanning of Native Chinese speakers was performed on a 3T Siemens MRI scanner in our laboratory in Beijing, China. Scanning of native English speakers was performed in London, UK in a previous study by Leff in 2008. In both scanning, thirty-five transversal slices of functional images were acquired using a gradient-echo echo-planar imaging (EPI) pulse sequence (TR/TE/θ = 2.08 s/30 ms/90º, 64 x 64 x 35 matrix with 3 x 3 x 3 mm^3 spatial resolution). The visual displays were presented through a LCD projector onto a rear-projection screen located over the

subject's head, viewed with an angled mirror positioning on the head-coil. The auditory stimuli were presented binaurally using a pair of home-made MRI compatible headphones that provided 25-30 dB/SPL attenuation of the scanner noise. In consideration of the scanner noise generated by EPI sequence, participants were required to adjust sound volume of the sample auditory stimuli to a clear-and-comfortable level during a pilot EPI scan before the experiment. After the volume adjustment, participants were instructed to pay attention to the experimental task. Four sessions of functional task scanning were acquired while participants performed gender judgment of the auditory stimuli. Each session started with a blank screen for 10 s, and followed by nine blocks of auditory stimuli, lasted 378.56 s in total. After the functional scanning, high resolution anatomical images were obtained using a 3D T_1-weighted MPRAGE sequence (TR/TE = 2.6 s/3.02 ms, 224 x 256 x 176 matrix with 1 x 1 x 1 mm^3 spatial resolution).

Behavioral Data Analysis. Response accuracy and reaction time were recorded for each type of stimuli. Since our interests in this research were focused on the processing of speech intelligibility, paired *t*-tests were conducted to compare the differences in behavioral performance between intelligible speech (averaging across idioms and rearranged idioms) and unintelligible speech (time-reversed idioms). The results of behavioral data in English participant were reported in a previous paper (21). The mean response accuracy of Chinese participants for gender judgment across all auditory stimuli was 98.5%. Paired *t*-test of response accuracy showed that subjects

made significantly more accurate gender judgment for intelligible speech than for time-reversed speech (t = 3.612, P < 0.01; intelligible: 99.0% ± 1.8%; time-reversed: 98.0% ± 2.3%), whereas there was no significant difference in reaction time between the two speech types (overall reaction time 1731 ± 176 ms).

fMRI Data Analysis. Statistical Parametric Mapping software (SPM8, Wellcome Trust Centre for Neuroimaging, UK) was used for imaging data processing and analysis. Native English speakers participated in the experiment in a previous study and the fMRI data were reanalyzed in the current research under the identical procedure. The EPI images were realigned to the first scan to correct head motion. Then, the mean image produced during the process of realignment and the realigned images were coregistered to the high-resolution T_1 anatomical image. All images were spatially normalized to standard Montreal Neurological Institute (MNI) space. The normalized EPI images were then spatially smoothed using an isotropic Gaussian kernel with a full-width at half maximum (FWHM) parameter for 8 mm. The functional imaging data were modeled using a box-car function with head motion parameters as unrelated regressors. Parameter estimates for each condition (three types of stimuli) were calculated from General Linear Model (GLM) based on hemodynamic response function with overall grand mean scaling. Whole-brain statistical parametric mapping analyses were performed and contrasts were then defined to reveal brain areas specifically involved in processing intelligible stimuli (Chinese idioms and rearranged idioms) and that of unintelligible stimuli

(time-reversed idioms). The t-contrast images were generated for comparison at each voxel. Statistical tests were first assessed in individual subjects, and random effect analyses were then conducted based on statistical parameter maps from each individual subject to allow population inference. A one-sample *t*-test was applied to determine group-level activation for intelligible effect.

To compare the neural activities of intelligibility effect between the Chinese group and the English group, parameter estimates of signal intensity for processing intelligible and unintelligible speech were extracted from regions of interests (ROIs) and compared using analysis of variance (ANOVA). The ROIs in the left anterior temporal cortex (region A), left posterior temporal cortex (region P), left inferior frontal cortex (region F) were defined as spheres with 6 mm diameter centered at the nearest peak voxels around the landmark ROIs' center observed in the intelligibility effect in both groups. Since right temporal pole only showed significant activation in the Chinese group, the landmark ROI definition of this region (R) in the English group were identical to the Chinese group.

Cross-sample analysis on activation. To examine the cross-sample difference on overall brain activation, we performed the whole-brain two sample t-test on the brain activation of intelligible effect between Chinese and English groups, but didn't find any significance in the bilateral temporal lobe as well as the left frontal lobe with $P<0.005$ voxel-level uncorrected threshold. Chinese group showed significantly stronger activation on the right occipital lobe and bilateral parietal lobe as Figure 1S

showed, which suggested a visual-spatial related processing during the comprehension of intelligible speech in Chinese. The absence of the significant statistical effect of rSTP between groups might be due to the smaller effect of rSTP comparing with the greater global activation such as on the left hemisphere (as shown in Figure 1*B*), which was based on parameter estimates of the activations by using whole-brain General Linear Model (GLM) estimation that employed the whole-brain average activation level as image scaling parameter. To fully eliminate the global effects on the rSTP activation, we conducted a local GLM estimation of the activation of both English and Chinese groups by using an explicit mask of the right temporal pole (Temporal_Pole_Sup_R and Temporal_Pole_Mid_R) from AAL template (33). Then we conducted a local two-sample t-test voxel-by-voxel in the area of the right temporal pole, and find a significant cluster (peak voxel at (40,4,-24), $P<0.05$ voxel-level uncorrected, cluster size=86, see the Figure S2 Left) that were more activated to intelligibility in Chinese than in English. The cortical inflation of the right hemisphere was performed with the Freesurfer image analysis suite (version 5.3), which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/) where the deep and light grey represented cortical sulcus and gyrus respectively. Note that the position of this cluster was slightly different with the one we found in Chinese group based on global GLM, which might be due to the difference of parameter estimation between the local and global GLM estimation.

Dynamic Causal Modeling (DCM) Analysis. After we identified the involvement of several brain regions in the processing of speech intelligibility, we conducted a dynamic causal modeling (DCM) analysis (20) to examine the effective connectivity among these brain regions. In DCM analysis, differential equations $\frac{dx}{dt} = (A + uB)x + Cu$ were employed to model the cortical dynamics of the neuronal populations in brain regions that were of interests, which describe how the current state of one neuronal population caused dynamics in another through synaptic connections that are intrinsic and fixed, and how these interactions change under the external influence like experimental manipulations or the influence of endogenous brain activity (34). Here, x is a vector representing the neural state of all brain regions that are in consideration, and u is also a vector representing all external input. Then, there are three matrices: A, B, and C. Matrix A represents the strength of fixed connection between the brain regions that are in consideration, in other words, the strength of connection when no external input exists. As it has been revealed that anatomical connection exists between each pair of the three brain regions (7, 18), it is assumed that reciprocal fixed connection between each pair of the brain regions existed, in other words, all parameters in matrix A would be set nonzero (34). Matrix B represents the strength of modulation of the connections by external inputs, i.e. the experimental manipulation, and in the current research, the processing of the speech intelligibility. Matrix C represents the direct influence to the brain regions by the external inputs that were generally convinced to be the sensory stimuli, such as auditory input in this experiment. Furthermore, it is also widely convinced that

modulatory stimuli, such as intelligibility of speech in this experiment, can only indirectly influence the brain regions, i.e., only some relative parameters in matrix B would be nonzero while all relative parameters in matrix C would be zero. In our model-space design, the supposition above was applied and all nonzero parameters in the matrixes were assumed Gaussian.

ROIs Selection and Time Series Extraction. In the current research, we were particularly interested in the brain regions in temporal and frontal cortex that showed significant involvement in the processing of speech intelligibility. The coordinates of the peak voxel in the clusters identified in the group-level random effect analysis of the intelligibility effect (comparing the neural activity during listening to intelligible speech vs. unintelligible speech, i.e., the time-reversed speech) with $P<0.05$ FWE corrected threshold were used to serve as a landmark for the individual ROIs. For each Chinese subject, ROIs were defined as 6mm-radium spheres centered at the peak activation voxel of intelligibility effect in the regions of our interests by searching voxels that survived at least $P<0.05$ threshold around the landmarks (revealed in group-level analysis) within 8 mm radium distance and within the same anatomical regions. Since the exact location of activation varied for each subject, this procedure ensured comparability of models as well as the extracted time series across subjects by applying both functional and anatomical constraints (35, 36). Given these criteria, we were able to define ROIs and extract time series for three-region model (P-A-F) in 22 out of 30 Chinese subjects, and for four-region model (P-A-F-R) in 18 out of 30

Chinese subjects, and the remaining subjects were excluded from the respective DCM analysis in whom the activation of at least one brain region failed to meet the criteria. For each ROI in individual subject, time series was extracted and computed as the first eigenvector across all suprathreshold voxels.

DCM Specification. We formed specific model space that contained the whole set of alternative models that were anatomically and functionally plausible, and employed Bayesian methods to estimate the model parameters. In model selection, considering the large number of the models to compare, we employed the family level inference and Bayesian model averaging within families on the model space (22) instead of the "single best model" selection strategy. This procedure provided inference about model parameters that could minimize the assumption bias on the model structure.

To establish the basic neural dynamic network of processing intelligible speech in Chinese and English, we first specified the model-space that consisted of three shared left hemisphere brain regions: A, P and F. For input "auditory", treated as a sensory input, seven alternate ways of how input "auditory" could enter the system were contained into the model space, that is, to pSTG only, to aSTG only, to pSTG and aSTG, to IFG only, to pSTG and IFG, to aSTG and IFG, to all three brain regions. For "intelligibility" treated as a modulatory input, it may modulate any combination of six directed connections among the three regions, results in 63 (i.e. 2^6 - 1) different model structures with the null model excluded from the analysis. Therefore, 63 different structure of modulatory input "intelligibility" crossed with the 7 different

structures of sensory input "auditory". A total of 441 models were compared in the three-region modeling.

To investigate the specific neural dynamic for the tonal language (Chinese), we then specified the four-region model-space that included right STP into the DCM analysis together with previously identified left pSTG, aSTG and IFG. Because of our prior results, we assumed that inputs were into left pSTG only. Modulatory input "intelligibility" was assumed to modulate any combination of the 12 directed connections among the four regions, with the null model excluded from the analysis. Thus, a total of 4095 (i.e. $2^{12}-1$) different models were estimated and compared.

Model Estimation and Bayesian Model Average (BMA). After the model space including all the candidate models were specified, all candidate models of all subjects were estimated using expectation maximization algorithm, calculating the parameters in each model as well as the free energy F as a good estimation of the log-evidence of each model. And the model estimation was compared among interested families.

Our first question was where sensory inputs entering the network. For the three-region model space, seven families with different auditory input were compared by random effect method (RFX) to determine the most possible input of the network and for four-region model the input was assume to be the same as that in three-region model. After the comparison of the model-input families, we then tested the existence of modulation on each connection. Models could be divided to paired-family based on whether the modulation to one connection is assumed to be nonzero, and then

compared by RFX method. The winning families were determined according to the exceedance probability of the RFX result with a high confidence and entered the BMA to generate a model summary that combined the likely parameters values for each family of good model fitness (22). The family-wise comparison for three-region model space showed the possibility of auditory input entering only pSTG was the highest (with an exceedance possibility of 0.53), thus indicating the auditory input entered the neural system exclusively from the posterior part of left temporal cortex, which was in consistent with the English speakers' data (21, 22). The family-wise comparison analysis for four-region model space showed that the modulations from pSTG to IFG, from IFG to aSTG, from IFG to rSTP, and from aSTG to rSTP were probably nonexistent (all with exceedance possibility less than 0.01). Thus, there were 255 models survived the restriction above, and entered BMA to calculate the group-level parameters (means and standard errors).

Second-level Analysis of Model Parameters. A one-sample *t*-test was employed to examine the statistical significance of modulation (nonzero parameter in matrix B) across each group of subjects based on individual BMA results with threshold of $P < 0.05$ (FDR corrected). Two-sample *t*-tests were then employed to compare the fixed connections and the modulations by speech intelligibility between Chinese and English groups.

Behavioral Dictation in Chinese. Chinese subjects who participated in the MRI

experiment and were eligible in the DCM analysis were contacted four months after their brain scanning for a "surprise" dictation test based on self-volunteer. Eight subjects (4 males, 4 females, average 24.0 year-old) participated in the dictation study where they were presented the same intelligible stimuli of Chinese idioms and required to write down the phrase they heard. The dictation results were classified into three categories: (1) Correct (90.6%±2.5%, mean±standard deviation), (2) Phonologically correct (8.7%±2.3%, where the word identity was wrong but the pronunciation was correct) and (3) Wrong (0.7%±0.5% , both the word identity and the pronunciation were wrong). The performance of the dictation was calculated as the percentage of proportional of their writing based on these three categories. We then conducted correlation analysis to investigate the correlations between the dictation results and the modulation strength of brain network connections from the DCM analysis (See Fig. 3*B* and Fig. S4).

References

1. Streeter LA (1976) Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature* 259:39-41.

2. Grossman T, Oberecker R, Koch ST, Friederici AD (2010) The developmental origins of voice processing in the human brain. *Neuron* 65:852-858.

3. Kuhl P, Rivera-Gaxiola M (2008) Neural substrates of langauge acquisition. *Annu Rev Neurosci* 31:511-534.

4. Stager CL, Werker JF (1997) Infants listen for more phonetic detail in speech

perception than in word-learning tasks. *Nature* 388:381-382.

5. Binder JR, et al. (2000) Human temporal lobe activation by speech and nonspeech sounds. *Cereb Cortex* 10:512-528.

6. Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* 343(6174):1006-1010.

7. Hickok G, Peoppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393-402.

8. Roth G, Dicke U (2005) Evolution of the brain and intelligence. *Trends Cogn Sci* 9(5):250-257.

9. Seidenberg MS (1997) Language acquisition and use: learning and applying prbablistic constraints. *Science* 275:1599-1603.

10. Cook VJ, Newson M (2007) *Chomsky's Universal Grammar: An Introduction* (Third Edition). Malden, MA; Oxford : Blackwell Pub.

11. Ueno T, Saito S, Rogers TT, Lambon Ralph MA (2011) Lichtheim 2: Synthesizing aphasia and the neural basis of langauge in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron* 72:385-396.

12. Ueno T, Lambon Ralph MA (2013) The roles of the "ventral"semantic and "dorsal" pathways in conduite d'approche: a neuroanatomically-constrained computational modeling investigation. *Front. Hum. Neurosci.* 7(422):1-7.

13. Scott SK, Blank CC, Rosen S, Wise RJS (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123:2400-2406.

14. Hickok G, Peoppel D (2004) Dorsal and ventral streams: a framework for

understanding aspects of the functional anatomy of language. *Cognition* 92:67-99.

15. Evans S, et al. (2014) The pathways for intelligible speech: multivariate and univariate perspectives. *Cereb Cortex* 24(9): 2350-2361.

16. Binder JF, et al. (2011) Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study. *Neuroimage* 54:1465-1475.

17. Kaas JH, Hackett TA (2000) *Proc Natl Acad Sci USA* 97(22):11793-11799.

18. Friederici AD (2011) The brain basis of language processing: from structure to function. *Physiol Rev* 91:1357-1392.

19. Ueno T et al. (2014) No lost in translation: generalization of the primary systems hypothesis to Japanese-specific language processes. *J. Cogn. Neurosci.* 26(2):433-446.

20. Friston KJ, Harrison L, Penny W (2003) Dynamic causal modlling. *Neuroimage* 19:1273-1302.

21. Leff AP, et al. (2008) The cortical dynamics of intelligible speech. *J Neurosci* 28(49):13209-13215.

22. Penny WD, et al. (2010) Comparing families of dynamic causal models. *PLoS Compu Biol* 6(3):e1000709.

23. Crinion JT, et al. (2009) Neuronantomical markers of speaking Chinese. *Hum Brain Mapp* 30:4108-4115.

24. Hannas WC (1996) *Asia's Orthographic Dilemma*. (University of Hawaii Press, Honolulu), pp181.

25. Visser M, Lambon Ralph MA (2011) Differential contributions of bilateral ventral

anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. *J Cogn Neurosci* 23(10):3121-3131.

26. Friston KJ (2005) A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* 360:815-836.

27. Zatorre P, Gandour J (2008) Neural specializations for speech and pitch: moving beyong the dichotomies. *Philos Trans R Soc Lond B Biol Sci* 363:1087-1104.

28. Obleser J, Wise RJ, Dresner M, Scott SK (2007) Functional integration across brain regions improves speech perception under adverse listening conditions. *J Neurosci* 27(9):2283-2289.

29. Price CJ (2012) A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage* 62:816-847.

30. Gandour J et al. (2004) Hemispheric roles in the perception of speech prosody. *Neuroimage* 23:344-357.

31. Binder JR, Desai RH (2011) The neurobiology of semantic memory. *Trends Cogn Sci* 15(11):527-536.

32. Cronk BC, Schweigert WA (1992) The comprehension of idioms: The effects of familiarity, literalness, and usage. *Appl Psycholinguist* 13: 131-146.

33. Tzourio-Mazoyer N et al. (2002) Automated anatomical labelling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single subject brain. *Neuroimage* 15: 273-289.

34. Stephan KE, et al. (2010) Ten simple rules for dynamic causal modeling. *Neuroimage* 49:3099-3109.

35. Stephan KE, Weiskopf N, Drysdale PM, Robinson PA, Friston KJ (2007) Comparing hemodynamic models with DCM. *Neuroimage* 38:387-401.

36. Heim S, et al. (2009) Effective connectivity of the left BA 44, BA 45, and inferior temporal gyrus during lexical and phonological decisions identified with DCM. *Hum Brain Mapp* 30:392-402.

Figure Legends

Fig. 1. Brain activations during the processing of intelligible speech. (*A*) Activations shown in the effect of intelligibility (intelligible speech > unintelligible speech) in native English speakers and native Chinese speakers. The brain areas that entered into DCM analysis were defined with a threshold of $P < 0.05$ FWE corrected, which are labeled with black arrows. (Left hemisphere: F=inferior frontal gyrus, A=anterior superior temporal gyrus, P=posterior middle/superior temporal gyrus; Right hemisphere: R=superior temporal pole/gyrus, Table S1; for display purposes, Figure 1A was shown in threshold P<0.005 voxel-level uncorrected and minimum cluster size 50 voxels); (*B*) ROI analysis for all four brain regions of interests, showing that the brain activity intensity in left pSTG(t_{54} = -0.423, P = 0.67), aSTG (t_{54} = -0.462, P = 0.65) and IFG (t_{54} = -0.275, P = 0.78) is compatible between the Chinese and English groups; (*C*) A local two-sample t-test between Chinese and English groups showed significant differences on right temporal pole activation of the intelligibility effect (Left) with threshold of $P < 0.05$ voxel-level uncorrected and minimum cluster size of 50 voxels, ROI analysis comparing the parameter estimates for signal intensity in region R (rSTP) between intelligible and unintelligible speech in native Chinese and native English speakers, showing a significant interaction effect ($P < 0.01$) for brain activation between the intelligibility of the speech and language group in rSTP.

Fig. 2. Results of the DCM BMA analysis. (*A*) Three-region models (pSTG-aSTG-IFG, i.e. P-A-F) of processing intelligible speech in tonal language (i.e.

Chinese, left) and non-tonal language (i.e. English, right), generated from 441 models in total (Table 1). Green arrow is for the connection significantly modulated by intelligible speech in both languages (left lower panel); the red arrows for connections significantly activated in one language compared to the other (middle and right lower panels); (*B*) Four-region model (pSTG-aSTG-IFG-rSTP, i.e. P-A-F-R) of processing intelligible speech for Mandarin Chinese speakers only, that included the right temporal pole in the dynamic network, generated from 4095 models in total (Table 2). The auditory stimuli entered the neural system via pSTG (P) in all models, and the arrowed lines display the connections showed significantly enhanced (solid) or decreased (dashed) modularity of speech intelligibility with average modulatory parameter estimates ± SEMs (s^{-1}) shown alongside ($P < 0.05$, FDR corrected).

Fig. 3. Correlations of modulation strength of brain connections in Chinese. (*A*) Positive correlation was found between the modulation strength on connections of left aSTG-to-pSTG and right aSTP-to-left pSTG. ($r = 0.896$, $P < 0.001$, see Fig. S3; after one outlier removed, $r = 0.465$, $P = 0.06$), no significant correlation of modulation strength was found in the cortical dynamics of English (all $P > 0.5$); (B) Modulation strength on connections of left- and right- aSTG to left inferior frontal gyrus predicted the individual behavior performance of idiom dictations on the percentage proportion of correctly identified word (left aSTG-to-IFG: three-region DCM analysis $r = -0.811$, $P = 0.015$ showed in figure, four-region DCM analysis $r = -0.703$, $P = 0.052$; right aSTG/STP-to-IFG: $r = -0.719$, $P = 0.045$).

Table 1. Parameter estimates (s^{-1}) of modulation of speech intelligibility on connections in the three-region models (P-A-F) of the Chinese and English groups

Connection	Tonal (*Chinese*)			Non-tonal (*English*)			*Chinese* vs. *English*
	mean	SEM	*t*	mean	SEM	*t*	*t*
From pSTG to							
aSTG	0.182	0.021	8.94**	0.093	0.026	3.63**	0.84
IFG	-0.083	0.019	-4.33**	0.177	0.026	6.78**	-2.66*
From aSTG to							
pSTG	0.254	0.026	9.79**	-0.097	0.027	-3.59**	2.18†
IFG	0.247	0.018	13.62**	-0.025	0.025	-1.01	2.76*
From IFG to							
pSTG	0.026	0.023	1.13	-0.078	0.027	-2.87*	1.57
aSTG	-0.100	0.017	-5.90**	0.008	0.026	0.31	-0.96

SEM = standard error.

* $P < 0.05$, FDR corrected.

** $P < 0.01$, FDR corrected.

† $P < 0.05$, uncorrected.

Table 2. Parameter estimates (s^{-1}) of modulation of speech intelligibility on connections in the four-region model (P-A-F-R) of the Chinese group

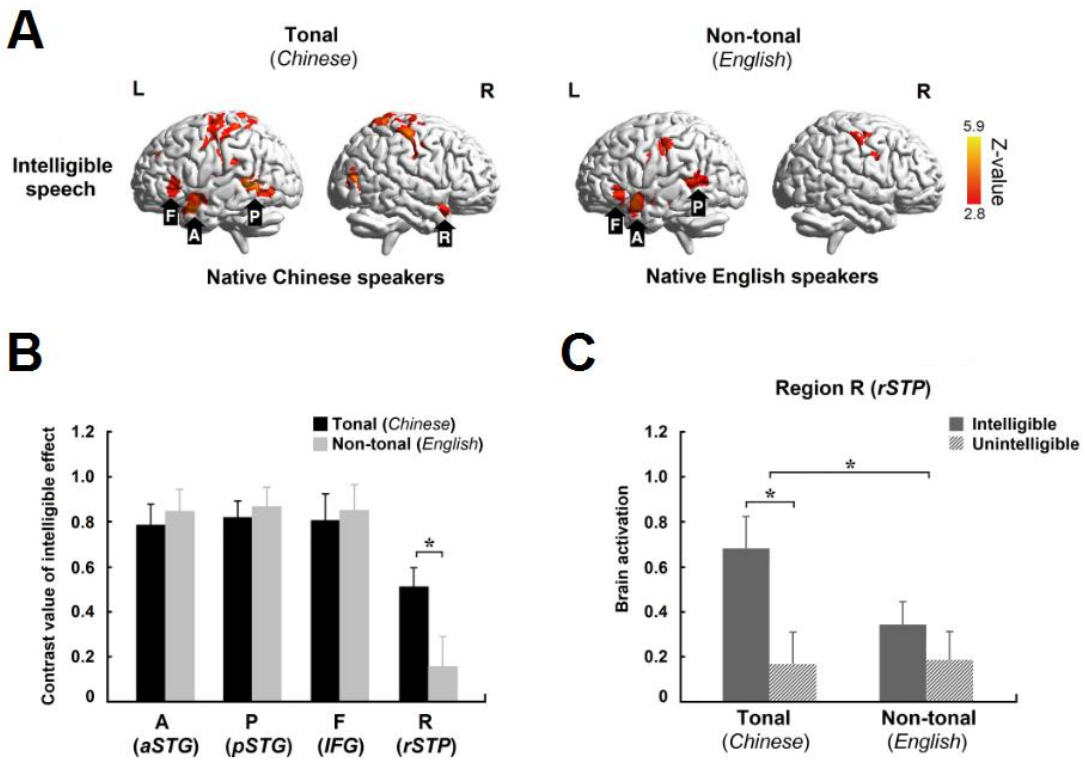
Connection	Intelligibility Modulation		
	mean	**SEM**	***t***
From pSTG to			
aSTG	0.052	0.020	2.59*
rSTP	0.165	0.018	9.03**
From aSTG to			
pSTG	0.190	0.029	6.54**
IFG	0.102	0.022	4.74**
From IFG to			
pSTG	0.022	0.030	0.72
From rSTP to			
pSTG	0.124	0.034	3.63**
aSTG	-0.192	0.029	-6.54**
IFG	0.142	0.025	5.69**

SEM = standard error.

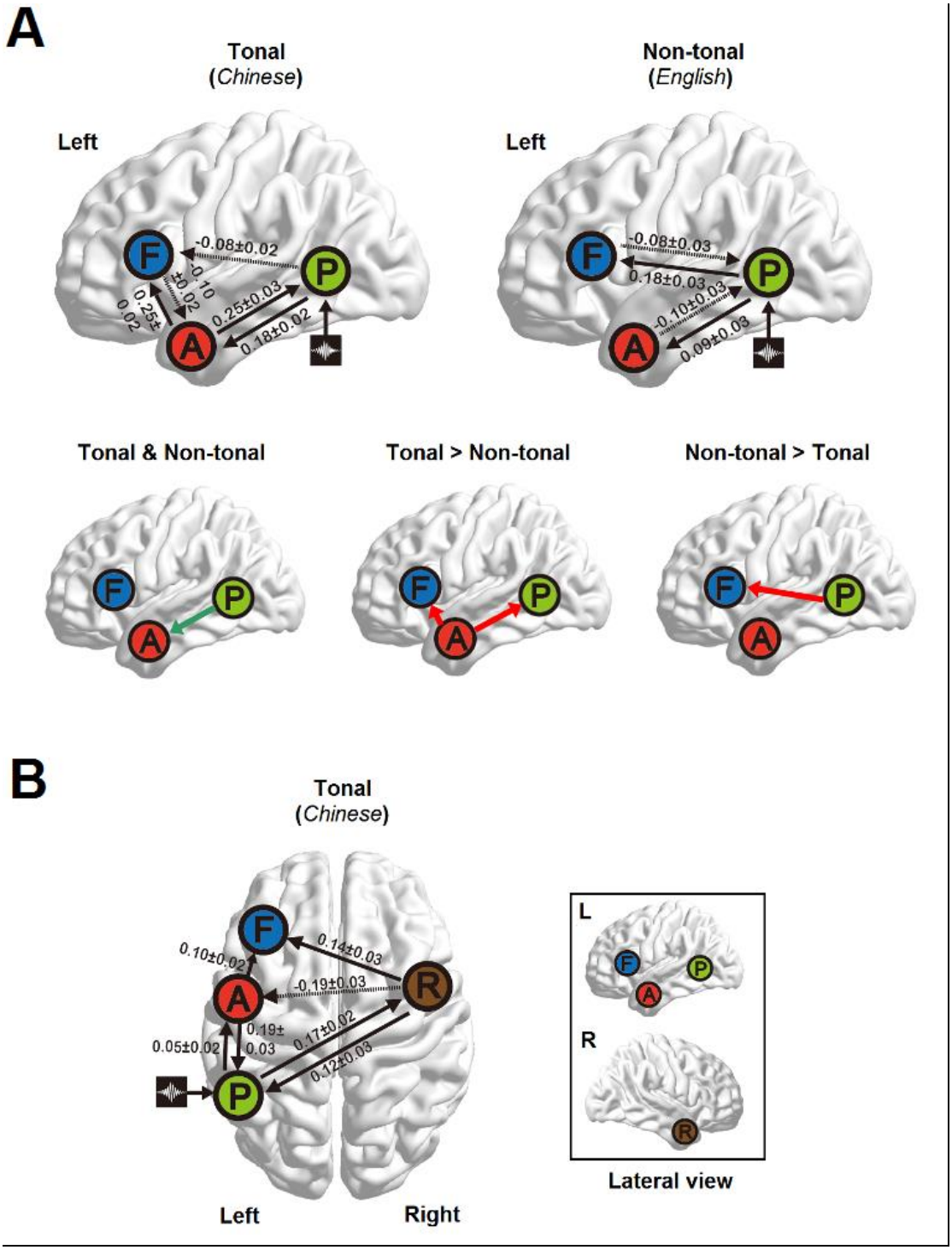
* $P < 0.05$, FDR corrected.

** $P < 0.01$, FDR corrected.

[Figure 1]


A
Tonal
(Chinese)
Non-tonal
(English)
L
R
L
R
Intelligible
speech
F
A
P
R
F
A
P
5.9
Z-value
2.8
Native Chinese speakers
Native English speakers
B
Tonal (Chinese)
Non-tonal (English)
Contrast value of intelligible effect
1.2
1.0
0.8
0.6
0.4
0.2
0
*
A
(aSTG)
P
(pSTG)
F
(IFG)
R
(rSTP)
C
Region R (rSTP)
Intelligible
Unintelligible
Brain activation
1.2
1.0
0.8
0.6
0.4
0.2
0
*
*
Tonal
(Chinese)
Non-tonal
(English)

[Figure 2]


A
Tonal
(Chinese)
Non-tonal
(English)
Left
Left
F
P
A
-0.08±0.02
-0.10
±0.02
0.25±
0.02
0.25±0.03
0.18±0.02
-0.08±0.03
0.18±0.03
-0.10±0.03
0.09±0.03
Tonal & Non-tonal
Tonal > Non-tonal
Non-tonal > Tonal
B
Tonal
(Chinese)
R
0.14±0.03
0.10±0.02
-0.19±0.03
0.19±
0.03
0.05±0.02
0.17±0.02
0.12±0.03
L
R
Lateral view
Left
Right

[Figure 3]

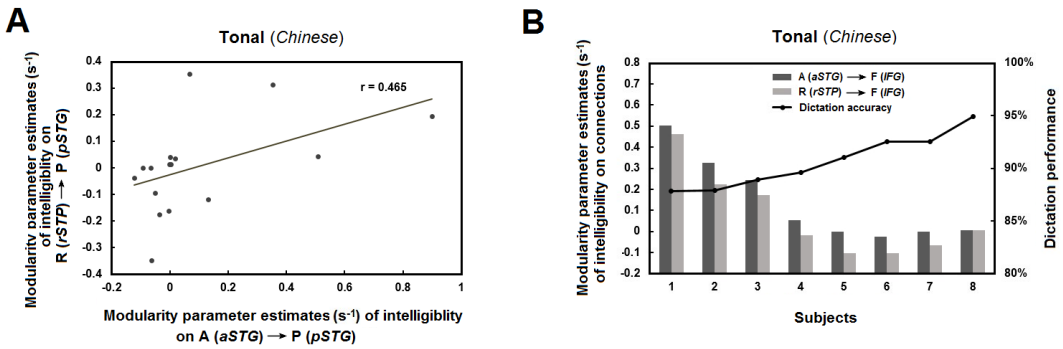

A
Tonal (Chinese)
r = 0.465
Modularity parameter estimates (s-1) of intelligiblity on R (rSTP) → P (pSTG)
Modularity parameter estimates (s-1) of intelligiblity on A (aSTG) → P (pSTG)
B
Tonal (Chinese)
A (aSTG) → F (IFG)
R (rSTP) → F (IFG)
Dictation accuracy
Modularity parameter estimates (s-1) of intelligibility on connections
Dictation performance
Subjects

Table S1. Brain activations of the Chinese group

Region	BA	MNI coordinates			Voxels	Z
		x	y	z		
Auditory Effect (I + C + TR > Rest)						
L. Superior Temporal Gyrus	22	-54	-14	6	1878	Inf
R. Superior Temporal Gyrus	22	60	-14	4	1858	Inf
Supplementary Motor Cortex	6	2	6	64	832	6.83
L. Lingual Gyrus	18	-18	-80	-14	423	6.68
L. Putamen	-	-22	8	2	507	6.59
R. Cerebelum Lobule VI	-	28	-56	-28	348	5.98
R. Putamen	-	24	8	4	219	5.92
L. Precentral Gyrus	6	-44	-2	58	305	5.87
Intelligibility Effect (I + C > TR)						
L. Superior Temporal Gyrus (A)	38	-50	14	-18	446	5.9
L. Posterior Middle Temporal Gyrus (P)	22	-60	-48	4	638	5.69
R. Superior Temporal Gyrus (R)	38	48	16	-18	243	3.23
L. Pars triangularis of Inferior Frontal Gyrus (F)	45	-48	30	-2	113	3.14

Brain activations in the Chinese group in Auditory contrast (Intelligible + Unintelligible > Baseline) and Region-of-Interests in intelligibility contrast (Intelligible > Unintelligible) that served as the landmark in individual ROI extractions, with a threshold of $P < 0.05$ FWE corrected (L.=left hemisphere, R.=right hemisphere). Brain activations in the English group have been reported in a previous study (12).

Figure S1. Whole-brain two-sample t-test on the brain activation of intelligible effect between Chinese and English groups showed no difference in the bilateral temporal lobe as well as the left frontal lobe.($P < 0.005$ voxel-level uncorrected)

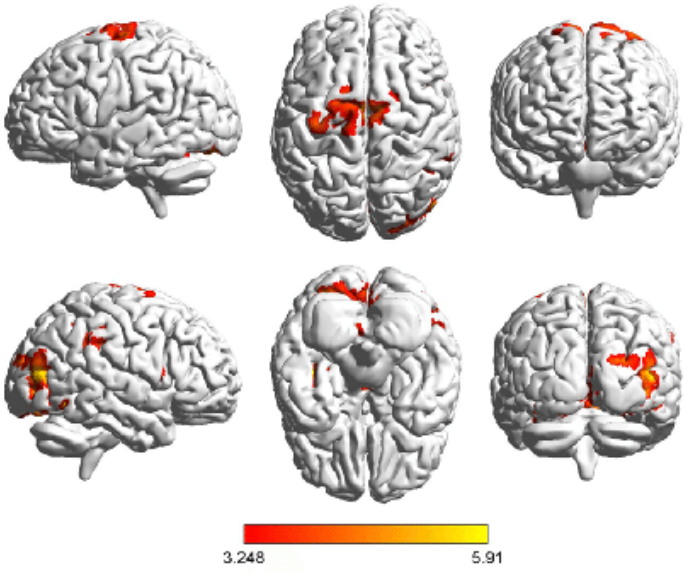


Figure S2. Regional two-sample t-test on the brain activation of intelligible effect between Chinese and English groups showed a significant cluster of 86 voxels that were activated stronger in Chinese than in English ($P < 0.05$ voxel-level uncorrected).

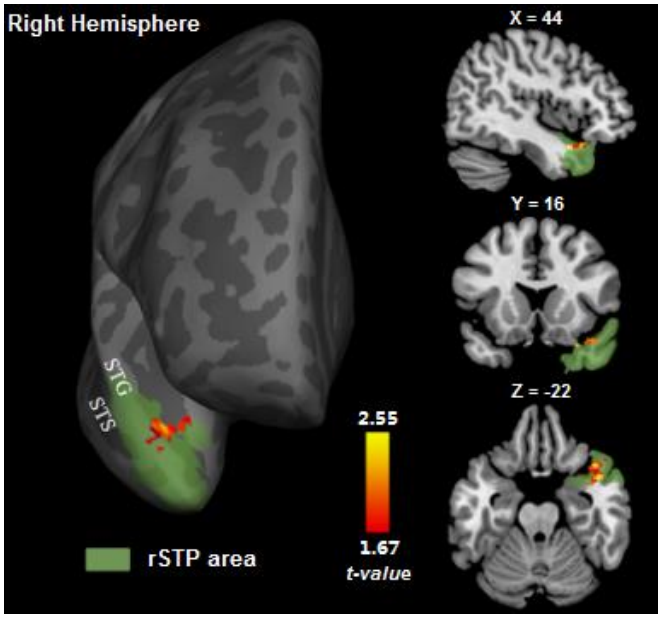


Figure S3. Correlation of modulation strength of intelligibility on connections of bilateral anterior temporal regions with left posterior region of superior temporal cortex in Chinese, before removing the outlier data (r = 0.896, P < 0.001)

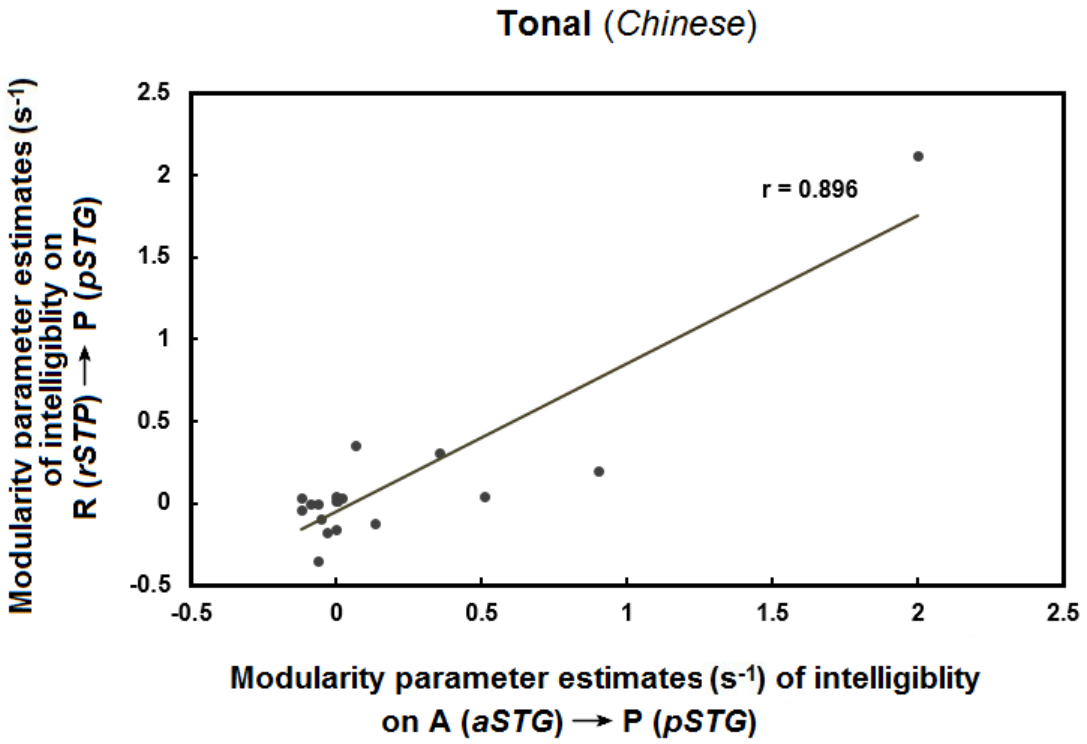


Figure S4. Correlations of modulation effects of connections and behavioral performance in idiom dictation. The individual modulation effects of intelligible speech on connections between bilateral anterior temporal regions (both left aSTG and right aSTG/STP) with left inferior frontal gyrus (IFG) region were negatively correlated with their overall accuracy ("Correct") during the dictation test (Fig. 3*B*), whereas it was positively correlated with individual's performance in phonologically correct portion ("Phonologically Correct", left aSTG-to-IFG: r = 0.723, P = 0.043; right aSTG/STP-to-left IFG: r = 0.681, P = 0.063). In the dictation test, three categories were classified based on individual's performance: (1) "Correct", where idioms were dictated correctly; (2) "Phonologically Correct", pronunciation of the idioms were shown to be dictated correctly, but the word were not correctly identified; (3) "Wrong", neither the pronunciation nor the word were identified correctly.